ORIGINAL RESEARCH PAPER

Ophthalomology

KEY WORDS: Large Language Models, Vision Large Language Models, Ophthalmology, Artificial Intelligence.

PERFORMANCE AND CLINICAL APPLICABILITY OF LARGE LANGUAGE MODELS IN OPHTHALMOLOGY

Dr. Yugindu Raizada* Post Graduate Student Ophthalmology (Batch 2022), ASCOMS & Hospital, Jammu *Corresponding Author

Dr. Reem Javeed Post Graduate Student Ophthalmology (Batch 2022), ASCOMS & Hospital, Jammu

Dr. Divyanshu Gupta Post Graduate Resident Ophthalmology (Batch 2023), ASCOMS & Hospital, Jammu

Dr. Renu Hashia Dhar Prof & Head, Department Of Ophthalmology, ASCOMS & Hospital, Jammu

ABSTRACT

This review synthesises current evidence on Large Language Models (LLMs) and Vision Large Language Models (VLLMs) in ophthalmology. LLMs show high accuracy in text-based tasks, with some models exceeding human performance in areas like glaucoma diagnosis. However, VLLMs exhibit a significant performance drop in image-based tasks without textual descriptions, and while abnormality detection is high, diagnostic description quality is often suboptimal. Common errors include misinterpreting data and overlooking clinical signs. Despite their potential, LLMs and VLLMs are not yet suitable for autonomous clinical decision-making, emphasising the ongoing need for human oversight due to issues like "hallucinations" and data limitations.

INTRODUCTION

The rapid advancements in Artificial Intelligence (AI), particularly Large Language Models (LLMs) and Vision Large Language Models (VLLMs), are revolutionising various fields, including medicine. LLMs show promise in healthcare for personalised consultations, clinical decision support, surgical planning, and telemedicine. Ophthalmology, with its abundant patient data, extensive literature, and diverse imaging, is an especially fertile ground for AI integration. Researchers are exploring how LLMs can leverage ophthalmic data to aid diagnosis and clinical decision-making.

Before widespread clinical adoption, rigorous evaluation of LLM performance and safety is crucial. This paper presents a secondary data analysis, a method akin to a systematic review, to synthesise existing evidence on LLM and VLLM performance in ophthalmology. This approach provides a broad overview, validates findings through triangulation, enhances reliability, and identifies data gaps, while also being time and cost-efficient compared to primary research.

The dynamic nature of LLMs, with frequent new models and non-linear performance variations, necessitates continuous synthesis of findings. A secondary analysis is vital for tracking this evolving landscape, identifying persistent challenges, and guiding future research and clinical applications in this rapidly changing field.

Various models like DeepSeek-R1, Gemini 2.0 Pro, OpenAI o1, OpenAI o3-mini, Copilot Pro, ChatGPT 3.5, Claude Pro, ChatGPT 4.0, and GPT-4 were studied.

METHODS

This secondary data analysis uses a comprehensive literature review approach, adhering to systematic review principles. The findings are from recent, peer-reviewed studies evaluating LLM and VLLM performance in ophthalmological diagnosis, clinical reasoning, and management. Data extraction focused on AI models, task types, methodologies, performance metrics (e.g., accuracy, sensitivity), error patterns, and human comparisons. The extracted information was then used to identify themes and contradictions, with quantitative data presented in comparative tables for clear interpretation.

Inclusion Criteria :

- Studies directly evaluating the performance of LLMs or VLLMs in ophthalmological tasks.
- Performance metrics include, but are not limited to accuracy, completeness, and reasoning ability.
- Ophthalmological tasks encompass diagnosis, management, and image interpretation.

Exclusion Criteria:

- Documents not directly related to the performance evaluation of LLMs/VLLMs in ophthalmology.
- Discussions on general AI concepts without specific ophthalmological application.

RESULTS

Large Language Models (LLMs) demonstrate significant prowess in text-based ophthalmology tasks. **DeepSeek-R1** showed superior performance in bilingual (Chinese and English) complex ophthalmology multiple-choice questions (MCQs), achieving 86.2% accuracy in Chinese and 80.8% in English. This was statistically significantly higher than other models like **Gemini 2.0 Pro**, **OpenAI o1**, and **OpenAI o3-mini**. DeepSeek-R1 also excelled in Chinese management questions, attributed to its innovative training using Chain-of-Thought (CoT) data.

GPT-4 proved highly proficient, even outperforming human fellowship-trained glaucoma specialists in diagnostic accuracy and completeness from text-based questions and real patient cases (P<.001 for both). For retina cases, **GPT-4** matched specialists in accuracy (P=.17) and exceeded them in completeness (P=.005). This suggests GPT-4's ability to synthesise complex clinical data for assessments and plans on par with seasoned sub-specialists.

However, LLM performance varies across models and versions. In a study of 15 LLMs on challenging open-text ophthalmic cases, **Copilot Pro** achieved the highest score (35/60, or 58%), followed by ChatGPT 3.5 (30/60). Interestingly, **ChatGPT 4.0** (19/60) performed worse than its predecessor, **ChatGPT 3.5**, on these specific difficult cases, highlighting that newer versions are not universally superior. Paid versions generally performed better than free counterparts. This variability underscores the need for context-dependent model selection and potentially

specialised LLMs for different clinical needs. Despite these differences, the evaluated LLMs generally exhibited similar underlying reasoning logic, involving the identification of key clues, analysis of clinical signs, and systematic elimination of alternative diagnoses.

Multimodal Performance And Image Interpretation

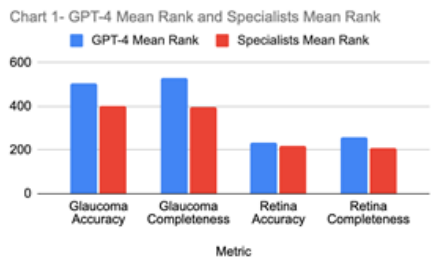
A significant limitation emerged in multimodal tasks involving complex ophthalmic images without explicit textual descriptions. GPT-4's diagnostic accuracy dropped from 47.9% (text-only prompts) to 38.4% when relying solely on figures, indicating a critical dependency on human-provided textual context rather than autonomous visual reasoning. Adding figure descriptions restored GPT-4's accuracy to near text-only levels (49.3%), further emphasising this reliance.

General-purpose Vision Large Language Models (VLLMs) showed high sensitivity in detecting general abnormalities from retinal photographs (e.g., GPT-4V default mode: 97.1%), but their diagnostic description quality was suboptimal, with only 21.2% of GPT-4V default mode responses rated as 'good'. Google Gemini exhibited lower abnormality detection sensitivity (41.1%) and also provided poor quality descriptions (28.6% 'good'). This suggests VLLMs struggle to translate raw visual data into nuanced, clinically meaningful diagnoses.

VLLMs also encountered difficulties with specific case types or image characteristics, such as Pathology and Tumors and Pediatric Ophthalmology cases without descriptions. Challenges included distinguishing closely spaced images, contextual mixing, and accurately interpreting composite images or laterality. Notably, all VLLMs, including Google Gemini, performed poorly in diagnosing Visually Significant Cataract from retinal photos, with Google Gemini misidentifying all such cases as normal.

Table 1: GPT-4 vs. Human Specialists

Metric	GPT-4 Mean Rank	Specialists Mean Rank
Glaucoma Accuracy	506.2	403.4
Glaucoma Completeness	528.3	398.7
Retina Accuracy	235.3	216.1
Retina Completeness	258.3	208.7



Limitations

Hallucinations And Reliability: A significant limitation is the propensity of LLMs to "hallucinate" or fabricate inaccurate information, often presented with high confidence. This necessitates rigorous fact-checking by clinicians for every LLM output, as their confidence is a function of probabilistic output, not factual truth. Google Gemini, for instance, exhibited a "misleading confident tone" despite poor output quality, posing a critical safety concern due to this "confidence-accuracy discrepancy." This reinforces that LLMs are assistive tools, not autonomous decision-makers, and unverified reliance could lead to misdiagnosis or inappropriate management.

Bias And Ethical Considerations: LLMs inherit biases from their training data. Furthermore, "designer guardrails" can lead models (e.g., Meta AI, Gemini) to refuse certain clinical questions, which, while intended for safety, can limit practical utility. Broader ethical concerns include responsibility, liability, and the lack of comprehensive regulatory oversight, highlighting the need for interdisciplinary collaboration to establish clear ethical guidelines and legal frameworks for clinical AI use.

Data Limitations And Knowledge Cutoff: A persistent issue is the inability to guarantee that test questions weren't part of the models' training data, leading to potential data leakage and inflated performance. Additionally, LLMs are inherently "dated" due to historical training datasets, meaning they cannot incorporate new medical advancements or current guidelines, impacting the relevance and timeliness of their responses.

Clinical Applicability And Future Directions

LLMs As Assistive Tools, Not Replacements: The consensus is that LLMs are not yet reliable enough for standalone patient care in ophthalmology. Their primary value lies as powerful assistive tools, augmenting human expertise, preventing oversights, and serving as diagnostic adjuncts, thereby expanding access to specialised medical knowledge. This advocates for a "human-AI teaming" paradigm, leveraging AI's strengths while mitigating weaknesses through paramount human oversight.

Future Research

Key areas include larger-scale, multicentered, longitudinal evaluations; quantifying LLM reasoning and enhancing interpretability; focusing on adductive reasoning common in medical diagnostics; exploring advanced multimodal strategies; and addressing data scalability and diversity for robust model training.

CONCLUSION

In summary, while Large Language Models demonstrate considerable promise in text-based ophthalmology tasks, their performance in multimodal (image-based) interpretation remains limited, often relying heavily on textual context. This review highlights common reasoning errors and the varying capabilities across different models and versions, underscoring that LLM performance is not uniformly correlated with model version or cost.

Currently, LLMs and VLLMs are valuable as assistive tools for ophthalmologists, enhancing diagnostic support and potentially preventing oversights. However, their current accuracy and reliability are not sufficient for autonomous clinical decision-making. The "confidence-accuracy discrepancy" observed in LLMs, coupled with their propensity for hallucinations and biases, necessitates continuous human oversight and rigorous fact-checking.

The future of AI in ophthalmology points towards a "human-AI teaming" paradigm, where AI augments human expertise rather than replacing it. This necessitates continued domain-specific fine-tuning, advanced prompting strategies, and robust validation studies. The ethical and regulatory landscape is currently lagging behind technological advancements, highlighting a critical need for interdisciplinary collaboration to establish clear guidelines. Ultimately, ongoing research and responsible integration are essential to harness the full potential of these powerful AI technologies for improved patient care.

REFERENCES:

[1] Huang, A. S., Hirabayashi, K., Barna, L., Parikh, D., & Pasquale, L. R. (2024). Assessment of a Large Language Model's Responses to Questions and Cases About Glaucoma and Retina Management. JAMA Ophthalmology. CorpusId: 267779484.

[2] Srinivasan, S., Ji, H., Chen, D., Wong, W., Soh, Z., Goh, J. H. L., ... Tham, Y. (2025). Can off-the-shelf visual large language models detect and diagnose ocular diseases from retinal photographs? BMJ Open Ophthalmology, 10. CorpusId:27762373.

[3] Mikhail, D., Milad, D., Antaki, F., Milad, J., Farah, A., Khairy, T., ... Duval, R. (2025). Multimodal Performance of GPT-4 in Complex Ophthalmology Cases. Journal of Personalized Medicine, 15. CorpusId:277999900

[4] Shanmugam, S., & Browning, D. J. (2024). Comparison of Large Language

Models in Diagnosis and Management of Challenging Clinical Cases. Clinical Ophthalmology (Auckland, N.Z.), 18, 3239–3247.

[5] Xu, P., Wu, Y., Jin, K., Chen, X., He, M., & Shi, D. (2025). DeepSeek-R1 Outperforms Gemini 2.0 Pro, OpenAI o1, and o3-mini in Bilingual Complex Ophthalmology Reasoning. ArXiv, abs/2502.17947.